Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2015, *9*, 2911-2919 2911

Open Access

Tibetan-Chinese Cross Language Text Similarity Calculation Based on LDA Topic Model

Sun Yuan[1,2,*] and Zhao Qian[1,2]

[1]*School of Information Engineering, Minzu University of China, Beijing, 100081, P.R. China*

[2]*Minority Languages Branch, National Language Resource and Monitoring Research Center*

Abstract: Topic model building is the basis and the most critical module of cross-language topic detection and tracking. Topic model also can be applied to cross-language text similarity calculation. It can improve the efficiency and the speed of calculation by reducing the texts' dimensionality. In this paper, we use the LDA model in cross-language text similarity computation to obtain Tibetan-Chinese comparable corpora: (1) Extending Tibetan-Chinese dictionary by extracting Tibetan-Chinese entities from Wikipedia. (2) Using topic model to make the texts mapped to the feature space of topics. (3) Calculating the similarity of two texts in different language according to the characteristics of the news text. The method for text similarity calculation based on LDA model reduces the dimensions of text space vector, and enhances the understanding of the text's semantics. It also improves the speed and efficiency of calculation.

Keywords: Cross-language, text similarity, comparable corpus, LDA model.

1. INTRODUCTION

Cross-language text similarity calculation has great significance in building parallel corpora and comparable corpora. It can provide the basis for cross-language information retrieval and machine translation. It is also the foundation of cross-language topic detection and tracking. When we calaulate the similarity of a Tibetan text and a Chinese text, we need a Tibetan-Chinese dictionary. But the construction of Tibetan-Chinese dictionary is not perfect and the development of Tibetan-Chinese machine translation technology is not mature. So we use the method based on Wikipedia to extract Tibetan-Chinese named entity for extending the dictionary.

Wikipedia is a free content, editable and multilingual online encyclopedia project. Expressions of the same entity in different languages are not parallel translation, but have a certain degree of similarity in content and form of the text. So Wikipedia has great value as the source of Tibetan-Chinese named entity extraction. Some researchers have discovered the value of Wikipedia [1] as the corpus of language information processing, and did some related research. Pablo Gamallo Otero [2] used Wikipedia's dual-language classification information to build comparable corpus automatically. He divided them into three kinds of corpora: non-alignment, strong alignment, weak alignment, according to the difference of the same entity's classified link information in two languages. Silviu Cucerzan [3] used Wikipedia for large scale named entities recognition and disambiguation. Torsten Zesch [4] extracted lexical semantic knowledge from Wikipedia, and listed advantages of language knowledge base based on collaborative encyclopedia website in building costs, timeliness, scale, quality and other aspects compared to traditional knowledge base.

The study of Tibetan-Chinese cross-language text similarity calculation is a necessary condition for building bilingual corpora, and it provides support for the cross-language research. Cross-language text similarity computing is mainly based on parallel corpora, multilingual dictionaries, and machine translation. Hasan [5] used mutual information and RIDF (Residual Inverse Document Frequency) to align Chinese-Japanese bilingual documents. Steinberger [6] used multilingual dictionary to translate multilingual texts into interlingua semantics, and then calculated the similarity of texts. Potthast [7] proposed CL-ESA algorithm (Cross-Language Explicit Semantic Analysis) based on parallel corpora to calculate the cross-language texts' similarity. Uszkoreit [8] identified bilingual aligned texts by querying the N-gram of bilingual texts based on bilingual dictionary. Wang Hongjun [9] used the improved Dice method based on statistical translation model and bilingual dictionary to compute text similarity.

In these methods mentioned above, calculating the similarity at the vocabulary frequency level will result in high time complexity. In order to solve these problems, some scholars try to match similar documents at the topic level. In recent years, the implied topic model is proposed in the field of topic detection and tracking. Topic model can automatically discover the topics from large document collection, and the topic is expressed as the probability distribution of vocabulary. It is widely used in areas such as the topic found and evolution analysis of academic literature

1874-110X/15

[10-11]. Preiss [12] analyzed topics of documents in different languages, and then translated the topic model in source language into the target language. The disadvantage of this method is that the quality of comparable corpora depends on topic model largely. Bilingual topic model generated from parallel data or multilingual aligned documents. The basic idea of this kind method is that documents in different languages share the same topic and each topic represented by words in different languages [13, 14].

The origin of topic model is latent semantic indexing (LSI) [15]. Latent semantic indexing is not a probability model, but its basic idea laid the basis for the development of the topic model. Based on the LSI, Hofmann [16] proposed probabilistic latent semantic indexing (pLSI), this model is considered a real topic model. Blei [17] proposed LDA (Latent Dirichlet Allocation) on the basis of pLSI extends to get a more complete probability generation model. LDA (Latent Dirichlet Allocation) model [9] is a kind of document topic generation model, also known as the three-layer Bayesian probability model. It contains three-layer structure: word, topic and document. Document to topic is the polynomial distribution, and topic to word is also polynomial distribution [18, 19]. It is an unsupervised machine learning method can be used to identify latent information from the large scale document collection or the corpora. It uses the bag of words which treats each document as a word vector. This converts text information for digital information which is easy modeling. This method does not consider the order between words, so it simplifies the complexity of the issue.

In the traditional topic model, topic collection is considered not change over time. But with the rapid development of the Internet, document collection is dynamic and has a strong timeliness. So some researchers proposed new topic model in that topic collection can change over time [20, 21]. The model can analyze the evolution of topic, and it can also be used in incident detection tasks [22]. In addition, because the number of documents rolls up over time, some researchers focused on online learning topic model task and put forward the efficient online learning algorithms: online variational inference [23], collapsed variational Bayes inference [24] and sparse stochastic inference [25]. These algorithms greatly improved learning efficiency of the topic model on the document flow data.

In the world of natural language processing, Tibetan resources are quite scarce compared to English and Chinese resources. A lot of basic research such as dictionaries, corpus, lexical analysis, syntax analysis, and named entity recognition are still not mature. Tibetan-Chinese bilingual information processing studies are still in their infancy, and current research on Tibetan language mainly in the area of lexical analysis, syntax analysis and machine translation. If we simply use vocabulary for large-scale Tibetan-Chinese text similarity calculation, this will lead to low efficiency and accuracy.

In this condition, this paper extracts Tibetan-Chinese entities and corpora from webs. In order to improve the efficiency of computation, we combined with vocabulary information at the topic level to calculate similarity. The main work of this paper is in the following.

1) Data acquisition contains Tibetan-Chinese dictionary construction and news corpora obtainment.

2) Using toipc model make text mapped to the space vector.

3) Matching Tibetan and Chinese texts by similarity calculation.

2. CROSS-LANGUAGE TEXT SIMILARITY CALCULATION BASED ON LDA

When establishing text space vector, the selection of key words has a great influence on text similarity calculation. Too many key words will cause the dimension of feature vector is too high and increase the complexity of calculation. On the contrary, too few key words will affect the correct match of two similar texts.

Considering the above two points, we adopt topic model for dimension reduction of text feature. Texts share implicit topic collection, and a topic is made up of some related key words according to a certain proportion. A text contains multiple topics, but it will focus on a topic mainly. We consider the topic whose proportion is the largest in all topics can represent the meaning of this text. So a text is expressed as a low dimensional vector. Through this method, dimension reduction can be implemented, at the same time key words can also reflect the main information of the text. Finally, efficiency of calculation can be improved.

The process of Tibetan and Chinese news text similarity calculation based on LDA is shown in Fig. (**1**).

3. DATA ACQUISITION

3.1. Tibetan-Chinese Dictionary Building

The Tibetan-Chinese dictionary which we build includes more than 50000 entries, extracting mainly from Wikipedia and a dictionary.

(1) Extract Tibetan-Chinese named entity from Wikipedia

Wikipedia is free and open, and all data can be downloaded, which are updated regularly. The editors introduce the entity in his language and cultural environment, so different language versions of the same entity are not directly parallel translation and have different perspectives and styles. It's convenient to extract named entities by Wikipedia's multilingual links.

The number of Tibetan entries is much less than Chinese entries, so we start from the Tibetan entity which has Chinese link to look for Chinese entity. For Tibetan entity which has Chinese link, we can obtain accurate Tibetan-Chinese entities by extracting the titles of Tibetan page and its Chinese links page as shown in Fig. (**2**).

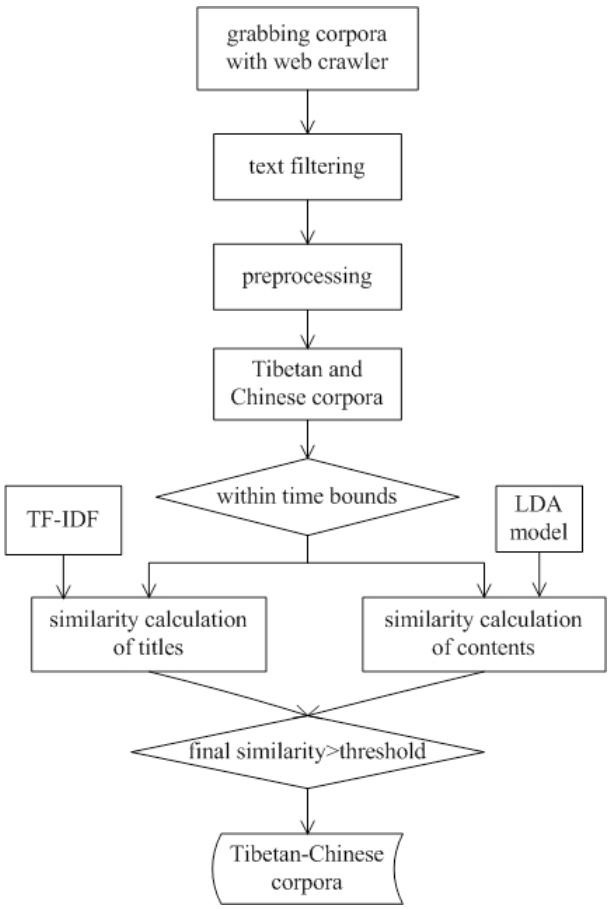

Fig. (1). The process of tibetan and chinese news text similarity calculation based on LDA.

Fig. (2). The page link of tibetan-chinese in wikipedia.

We use the web crawler to obtain Tibetan-Chinese entities by extracting the titles.

(2) Extract Tibetan-Chinese translation from Tibetan-Chinese dictionary

Due to the original Tibetan-Chinese dictionary contains a lot of redundant information, but we just need the corresponding entries in Tibetan and Chinese. So we extract the translation in accordance with the rules of words distribution.

3.2. Tibetan-Chinese News Corpora Acquisition

The construction of Tibetan corpus is lagging, this makes Tibetan corpus insufficient and the research on Tibetan-Chinese is more difficult. But we can get a large number of Tibetan corpora from the network.

We need to acquire Tibetan and Chinese news corpora by web crawlers from news site, which include headlines, time and content of news. Then filter out texts whose content is too little, so we obtain the initial bilingual corpora.

3.3. Data Preprocessing

Segmentation: We use the tool developed by National Language Resource and Monitoring Research Center for Tibetan word segmentation and ICTCLAS for Chinese word segmentation.

Removal of meaningless words: In Tibetan, case-auxiliary represents the relationship of words in a sentence. Case-auxiliary is a kind of function words which has no practical significance and refers to the difference between words. So we need remove case-auxiliary from Tibetan texts.

We also remove words which are not very helpful to the text but very high frequency according to Chinese stop-words.

Part of speech selection: Choose nouns and verbs which contain two words at least.

4. TOPIC MODELING

Generation process of LDA model can be expressed in Fig. (**3**).

K is the number of topics, M is the number of documents, N_m is the total number of words is the m document.

$\vec{\beta}$ is the Dirichlet prior parameter in the multinomial distribution of words under each topic. $\vec{\alpha}$ is the Dirichlet prior parameter in the multinomial distribution of topics under each document.

$z_{m,n}$ is the topic of the n word in the m document, $w_{m,n}$ is the n word of the m document.

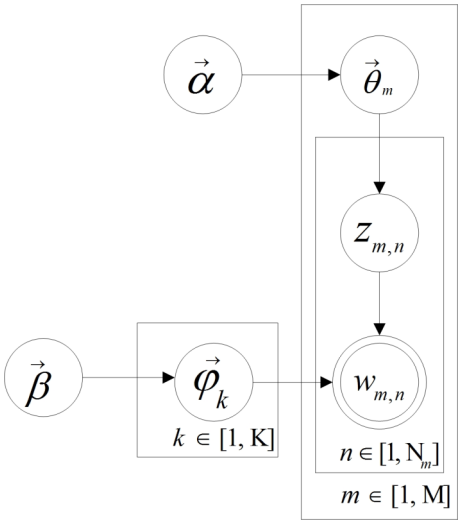

Fig. (3). Generation process of LDA model.

$\vec{\theta}_m$ is topic distribution of the m document. $\vec{\varphi}_k$ is word distribution of the k topic. The former is K dimensional vector (K is the number of topics), the latter is V dimensional vector (V is the number of words).

For a collection of documents, $w_{m,n}$ is the known variable which can be observed. $\vec{\alpha}$ and $\vec{\beta}$ are given prior parameters based on experience. $z_{m,n}$, $\vec{\theta}_m$ and $\vec{\varphi}_k$ are unknown latent variables, they are estimated according to the observed variables.

We estimate parameters of the LDA model by Gibbs sampling. The process is shown in Fig. (**4**).

First, the model assigns a topic $z^{(0)}$ to each word of the texts randomly, and then counts the number of the word v appears in each topic z and the number of words which are from each document m appears in the topic z.

Calculating $p(z_i | z_{-i}, d, w)$ in every round, which takes out the current distribution of topic words, in accordance with the topic of every other word distribution to estimate probability assigned to each topic of the current word. After we get the probability distribution of current word belongs to all topics, sample a new topic for this word according to the probability distribution. And then update the topic of the next word in the same way constantly until the distribution of topics and words no longer change. The final output are $\vec{\theta}_m$, $\vec{\varphi}_k$ and $z_{m,n}$.

$\alpha = 50 / K$, $\beta = 0.01$. Produce the topic-word distribution $\varphi_{k,v}$ according to equation (1).

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v} \tag{1}$$

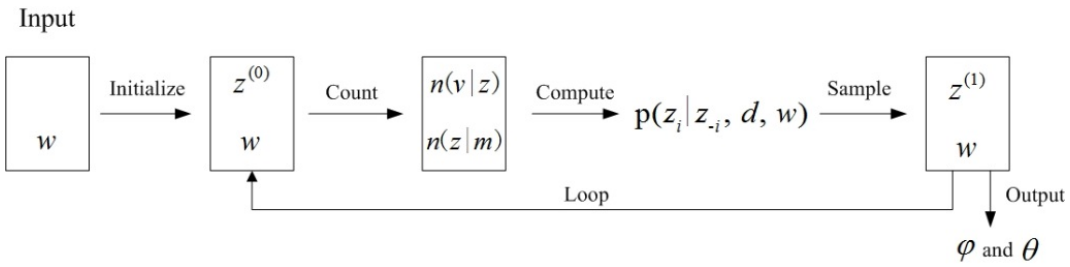

Fig. (4). The process of gibbs sampling.

For each document, produce the document-topic distribution $\theta_{m,k}$ according to equation (2).

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{2}$$

5. CROSS-LANGUAGE TEXT SIMILARITY CALCULATION

For network news, headline plays a decisive role. So it should have higher contribution to the final similarity. We calculate Tibetan and Chinese news text similarity using equation (3).

$$Sim(T_i, C_j) = ë\ ?\ Sim_{title}(T_i, C_j) + (1 - ë) \times Sim_{text}(T_i, C_j) \tag{3}$$

T_i represents Tibetan news text, C_j represents Chinese news text.

$Sim(T_i, C_j)$ is the similarity of T_i and C_j.

$Sim_{title}(T_i, C_j)$ is the similarity of headlines. $Sim_{text}(T_i, C_j)$ is the similarity of contents.

λ is an artificial selected parameter.

(1) Feature Selection

The strings match directly. There are a lot of texts involve time, digital or other strings in a large number of news text. Choosing these strings as characteristics and adding them to the feature vector of the text can increase the cross-language text matching degree.

Word pairs in translation. The Tibetan-Chinese word pairs by matching the Tibetan-Chinese dictionary.

After the feature selection, titles and contents of Tibetan and Chinese news text are represented respectively by the vector space.

$$T_i = (tw_1, tw_2, \cdots, tw_x)$$

$$C_j = (cw_1, cw_2, \cdots, cw_y)$$

(2) Similarity Calculation of Titles

Comparing to the content, the number of words in a title is small. So we can regard all words as the vector features after removing the meaningless words. We calculate TF-IDF as the weights of features. TF is the times of a word appear in a document. IDF is calculated as equation (4).

$$IDF = \log(\frac{D}{1 + D_w}) \tag{4}$$

D is the number of documents in corpora, D_w is the number of documents contain word w.

(3) Similarity Calculation of Contents

Calculate document-word distribution $\phi_{m,v}$ by equation (5).

$$\phi_{m,v} = \theta_{m,k} \times \varphi_{k,v} \tag{5}$$

Calculate similarity by cosine measure, as in equation (6).

$$cos(T_i, C_j) = \frac{\sum_{k=1}^{min(x,y)} tw_k \times cw_k}{\sqrt{\sum_{k=1}^{x} tw_k^2}\sqrt{\sum_{k=1}^{y} cw_k^2}} \tag{6}$$

6. EXPERIMENTAL RESULTS AND ANALYSIS

6.1. Experimental Data

The experimental data is from a real network, we use the network crawler to get news text from Chinese Version (http://tibet.news.cn/) and Tibetan Version (http://xizang.news.cn/) of Xinhua net's Tibet channel. After filter out texts whose content is less, we choose 1719 Chinese News texts and 677 Tibetan news texts in 2014. The information of experimental data is shown in Table **1**.

The evaluation indicator is precision P as in equation (7).

$$P = \frac{C_{correct}}{C_{extract}} \tag{7}$$

$C_{correct}$ is the number of cross-language similar documents which is correct matching. $C_{extract}$ is the number of documents which is matching.

6.2. The Effects of Topic Quantity Changes on the Result of the Experiment

Set the number of iterations to 1000. The number of topics K is set from 400 to 1200, with 100 increments.

Table 1. The information of original data.

Corpus category	Number of Texts	Number of Titles' Words	Number of Contents' Words	The Average Number of Words in a Text
Chinese Corpora	1719	32,082	1,309,040	761
Tibetan Corpora	677	19,791	274,573	405

Table 2. The precision under different number of topics.

K	P
400	0.67
500	0.71
600	0.72
700	0.70
800	0.66
900	0.66
1000	0.62
1100	0.63
1200	0.60

Table **2** shows the precision under different number of topics. And from Fig. (**5**), we can know the optimal number of topics is 600. On this occasion, we can get better performance. But within the scope of the test, the influence of K to the matching algorithm is not very significant. So the algorithm shows strong stability.

6.3. The Influence on Experimental Result of Iterations

Set the number of topic to 600. The number of iterations is set from 100 to 900 with 100 increments. The precision and time under different number of iterations is shown in Table **3**.

Fig. (**6**) shows that, with the increase of iterations, run time of programs is also on the rise. So we should balance the cost between time and accuracy. When the number of iterations is more than 500 times, accuracy has not improved significantly, and the time cost is on the contrary. Therefore, 500 times is relatively good choice.

Table 3. The precision and time under different number of iterations.

Iterations	P	Time(s)
100	0.59	430
200	0.61	912
300	0.65	1241
400	0.68	1689
500	0.70	2049
600	0.70	2673
700	0.71	2853
800	0.71	3460
900	0.71	3668

6.4. The Setting of λ

When we calculate similarity, the parameter λ has an important influence on the matching accuracy. The title holds the pivotal role in the news, so the choice of λ should

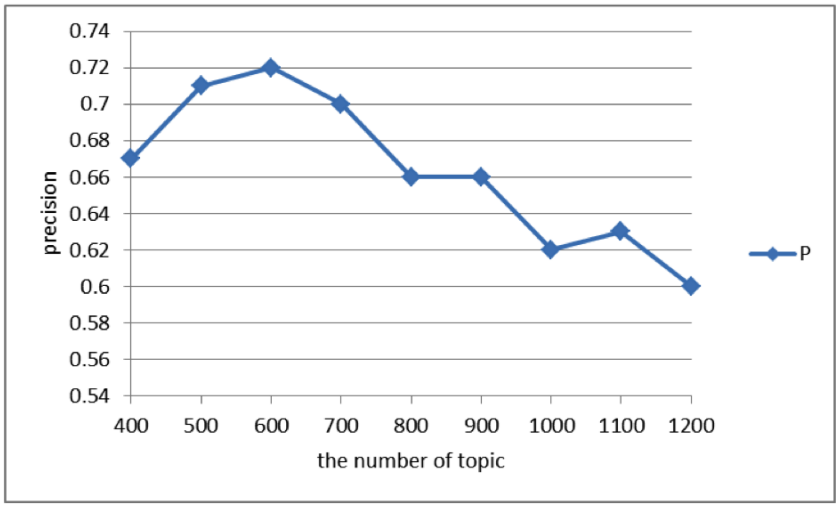

Fig. (5). The influence of the number of topic.

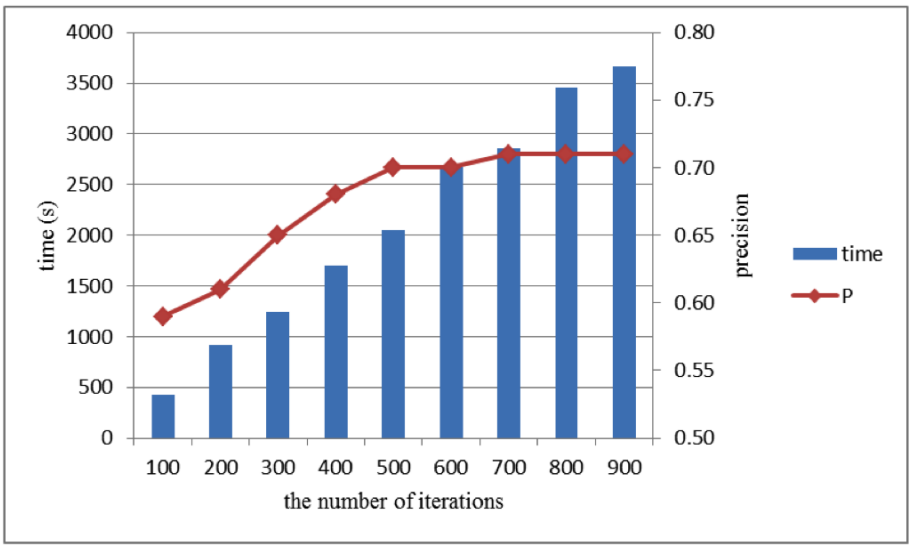

Fig. (6). The comparison under different number of iterations.

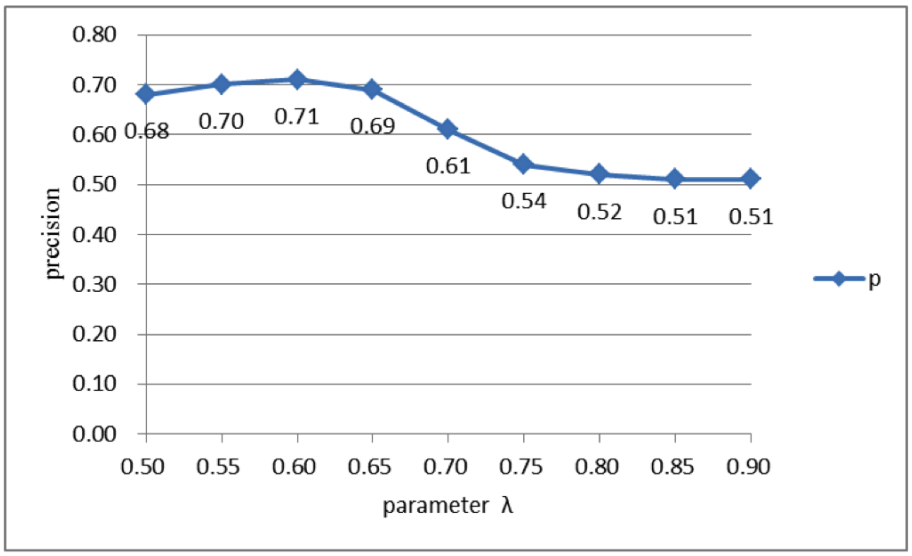

Fig. (7). The influence of the parameter λ.

show the importance of the title, meanwhile do not break the rationality at the same time.

According to Fig. (**7**), we can get relatively high accuracy when the parameter λ takes 0.6. When we take the figures larger than 0.6, the accuracy gradually drops. Because the number of words belong to the title is limited, if the title has a high weight, it will have a negative impact on the overall similarity.

6.5. The Choice in the Number of Topic Words

When the number of topic words number increases, the accuracy and the time of matching a Tibetan text and a Chinese text are both on the rise. Topic words chosen more, accurate rate will be higher, but the cost of time will be greater. Therefore, we must find a balance point, which makes the time cost in the acceptable range and the accuracy rate as high as possible. The precision and average time under different number of topic words is shown in Table **4**.

Table 4. The precision and average time under different number of topic words.

Topic Words	P	Average Time(s)
10	0.57	60
15	0.57	61.9
20	0.63	63.2
25	0.64	64
30	0.69	64.8
35	0.69	65.5
40	0.71	67.3
45	0.72	69.6
50	0.72	71

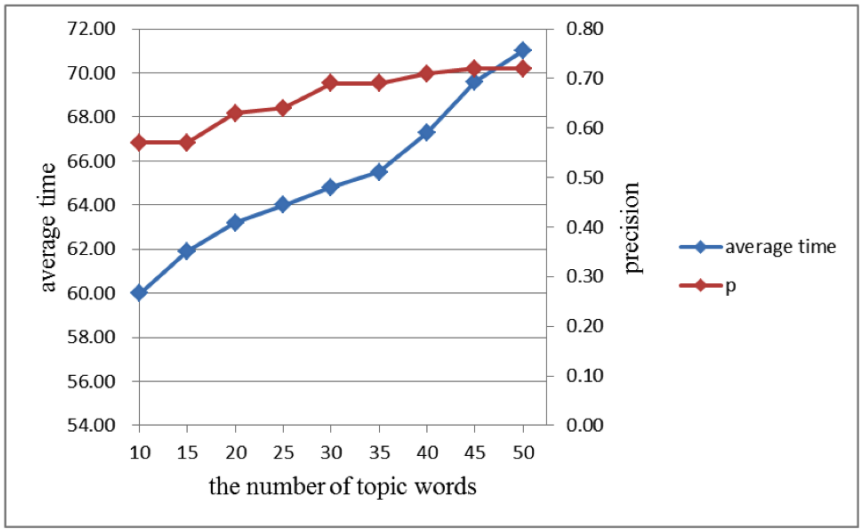

Fig. (8). The influence of the number of topic words.

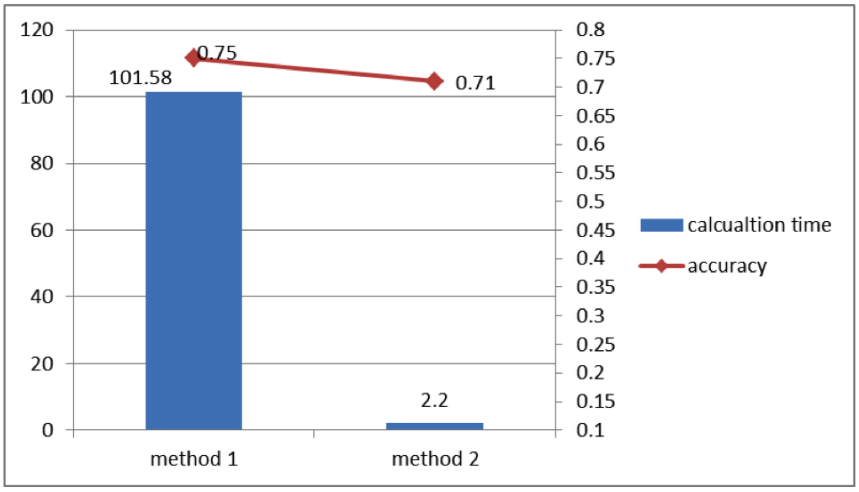

Fig. (9). The comparison of calculation time.

By the Fig. (**8**), we can see, when the number of topic words is 30, we can get acceptable time cost and good accuracy.

6.6. The Advantage of the LDA Model

As in Fig. (**9**), when K is set to 600, we compare the calculation time and accuracy of two methods. One method is the traditional method in word level; the other is used in this paper. In the case of accuracy is not dropped dramatically, the calculation time of two texts based on LDA is 2.2s on the average, in the case of not using the LDA model the time up to 101.58s on average. After LDA modeling, we achieve the dimension reduction of text space in contrast to the information of Table **1**. So this method based on LDA improves calculation efficiency.

CONCLUSION

In the cross-language text similarity calculation method based on LDA model, texts in the high dimensional space are mapped to low dimensional topic space. This method increases the calculation efficiency. But the main factors which affect correction of cross-language text matching are: the particle size of Tibetan and Chinese word segmentation is different, the accuracy of Tibetan word segmentation needs to be improved, and Tibetan-Chinese dictionary included unknown word. In future work, we will try to improve the LDA model, reduce the dependence on bilingual dictionary, and raise the accuracy of Tibetan word segmentation.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation (No. 61501529, 61331013), Beijing Higher Education Young Elite Teacher Project (No. YETP1291),

National Language Committee Project (No. YB125-139, ZDI125-36), and Minzu University of China "First Class University and First Discipline Project" (No. 10301-0150200518).

REFERENCES

[1] Wikipedia: About, http://en.wikipedia.org/wiki/Wikipedia:About

[2] P. G. Otero, and G. L. Isaac, "Wikipedia as multilingual source of comparable corpora ," In: *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, Malta, pp. 21-25, 2009.

[3] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data ," *EMNLP-Conll*, vol. 7, pp. 708-716, 2007.

[4] T. Zesch, C. Müller, and I. Gurevych, "Extracting lexical semantic knowledge from wikipedia and wiktionary," *LREC*, vol. 8, pp. 1646-1652, 2008.

[5] M. M Hasan, and Y. Matsumoto, "Multilingual document alignment-a study with chinese and japanese," *NLPRS*, pp.617-623, 2001.

[6] R. Steinberger, B. Pouliquen, and J. Hagman, "*Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc, Computational Linguistics and Intelligent Text Processing*," Springer Berlin Heidelberg, pp. 415-424, 2002.

[7] M. Potthast, B. Stein, and M. Anderka, "A wikipedia-based multilingual retrieval model," *Advances in Information Retrieval*, vol. 4956, 2008, pp. 522-530.

[8] J. Uszkoreit, J. M. Ponte, and A. C. Popat, "Large scale parallel document mining for machine translation," In: *Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics*, 2010, pp. 1101-1109.

[9] H. Wang, S. Shi, and S. Yu, "Cross-language similarity document retrieval, " *Journal of Chinese Information Processing*, vol. 21, no.1, pp. 30-37, 2007.

[10] D. M. Blei, and D. L. John, "A correlated topic model of science," *The Annals of Applied Statistics*, vol. 1, pp. 17-35, 2007.

[11] T. L. Griffiths, and M. Steyvers, "Finding scientific topics," In: *Proceedings of the National Academy of Sciences of the United States of America*, 2004, pp. 5228-5235.

[12] J. Preiss, "Identifying comparable corpora using LDA," In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*, Human Language Technologies. Association for Computational Linguistics, 2012, pp. 558-562.

[13] X. Ni, J. T. Sun, and J. Hu, "Mining multilingual topics from Wikipedia," In: *Proceedings of the 18th International Conference on World Wide Web,* ACM, 2009, pp. 1155-1156.

[14] D. Mimno, H. M. Wallach, and J. Naradowsky, "Polylingual topic models," In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, vol.2, pp. 880-889, 2009.

[15] S. C. Deerwester, S. T. Dumais, and T. K. Landauer, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no.6, pp. 391-407, 1990.

[16] T. Hofmann, "Probabilistic latent semantic indexing," In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1999, pp. 50-57.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[18] Z. Zhu, "*Research on Key Technology in Mining Web Bilingual Corpora*," University of Science and Technology of China, 2014.

[19] Q. Lu, "*Study on English-Chinese Cross-Language Topic Detection and Tracking Technology*," Minzu University of China, 2013.

[20] A. Ahmed, and P. X. Eric, "Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream," arXiv preprint arXiv, pp. 1203.3463, 2012.

[21] Y. Wang, E. Agichtein, and M. Benzi, "TM-LDA: efficient online modeling of latent topic transitions in social media," In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 123-13.

[22] G. Doyle, and E. Charles, "Accounting for burstiness in topic models," In: *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 281-288.

[23] M. D. Hoffman, M. B. David, and B. Francis, "Online learning for latent dirichlet allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856-864, 2010.

[24] I. Sato, K. Kenichi, and N. Hiroshi, "Deterministic single-pass algorithm for lda, " In: *Proceedings of the Conference on Advances in Neural Information Processing Systems(NIPS)*, 2010.

[25] D. Mimno, M. Hoffman and D. Blei, "*Sparse Stochastic Inference for Latent Dirichlet Allocation*," ICML, Edinburgh, Scotland, 2012.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015